

Towards Crowdsourcing and Cooperation in Linguistic Resources

Dmitry Ustalov^{1,2,3}

¹ Krasovsky Institute of Mathematics and Mechanics, Ekaterinburg, Russia,

² Ural Federal University, Ekaterinburg, Russia,

³ NLPub, Ekaterinburg, Russia

dau@imm.uran.ru

Abstract. Linguistic resources can be populated with data through the use of such approaches as crowdsourcing and gamification when motivated people are involved. However, current crowdsourcing genre taxonomies lack the concept of cooperation, which is the principal element of modern video games and may potentially drive the annotators' interest. This survey on crowdsourcing taxonomies and cooperation in linguistic resources provides recommendations on using cooperation in existent genres of crowdsourcing and an evidence of the efficiency of cooperation using a popular Russian linguistic resource created through crowdsourcing as an example.

Keywords: games with a purpose, mechanized labor, wisdom of the crowd, gamification, crowdsourcing, cooperation, linguistic resources.

1 Introduction

Crowdsourcing has become a mainstream and well-suited approach for solving many linguistic data gathering problems such as sense inventory creation [1], corpus annotation [2], information extraction [3], etc. However, its most effective use still remains a problem because human annotators' motivation and availability are tantalizingly constrained and it is crucial to get the most of performance from the effort interested people can make.

Another extremely popular term nowadays is *gamification*. The origin of the gamification concept is, of course, video game industry. The idea of gamification is in embedding interactive and game-based techniques into application to strengthen user engagement and increase the time spent annotating. Due to the insufficiency of exploration, gamification is more rarely used in academia when compared to the industry.

Cooperation is a major, if not principal, element of today's video games, which is confirmed by the presentations made in recent years at E3 — the largest video game exposition and event. Initially, multiplayer mode in video games was focused on *player versus player* competitions, but a few years ago the focus has changed to *cooperated human players versus AI* and *guild versus guild* games.

The work, as described in this paper, makes the following contributions: 1) it presents a survey on crowdsourcing taxonomies and cooperation in linguistic resources, 2) makes recommendations on using cooperation in existent genres of crowdsourcing, and 3) provides an evidence of the efficiency of cooperation represented by a popular Russian linguistic resource created through crowdsourcing.

The rest of this paper is organized as follows. Section 2 focuses on related work towards crowdsourcing genres and cooperation in linguistic resources. Section 3 is devoted to the cooperative aspect of crowdsourcing. Section 4 discusses cooperation using OpenCorpora as the example, which is a sufficiently popular Russian linguistic resource created through crowdsourcing. Section 5 interprets and explains the obtained results. Section 6 concludes with final remarks and directions for the future work.

2 Crowdsourcing Genres & Activities

Early studies on crowdsourcing genres in their wide definition were conducted in 2009. Quinn & Bederson in their technical report [4] proposed the term of *distributed human computation* along with the taxonomy of seven different genres of these computations such as games with a purpose, mechanized labor, wisdom of crowds, crowdsourcing, dual-purpose work, grand search, human-based genetic algorithms, and knowledge collection from volunteer contributions.

In the same year Yuen et al. also presented [5] another taxonomy of five crowdsourcing genres: initiatory human computation, distributed human computation, social game-based human computation with volunteers, paid engineers and online players, which is similar to the previously mentioned.

Many studies following the early ones are focused on classification of whether a crowdsourced project belongs to a specific class of the given taxonomy. For instance, Sabou et al. study of correlation between crowdsourcing genres [6], quality assessment [7], and guidelines on corpus annotation through crowdsourcing [2] align various best practices among the established genres.

There are other attempts to create a taxonomy of crowdsourcing genres. Zwass investigated the phenomena of *co-creation* [8] and proposed a taxonomy of user-created digital content which includes the following: knowledge compendia, consumer reviews, multimedia content, blogs, mashups, virtual worlds. The resulted taxonomy appears to be too general and, since it was not intended, does not fit the natural language processing field perfectly.

Erickson presented *four quadrant model* [9] composed of two orthogonal dichotomies to classify crowdsourcing projects: “same place–different places” and “same time–different times”. The resulted taxonomy tends to assign all the mentioned above crowdsourced projects to the “different places–different times” quadrant also called *Global Crowdsourcing*.

Some studies propose much narrower dichotomies. This is the case of the research conducted by Suendermann & Pieraccini [10], which introduces a concept of *private crowd* being a trade-off between two extremes: an inexpensive, highly available yet uncontrolled *public crowd* such as the Amazon’s one, and the

expensive to hire, high-quality and professional expert annotators. The *private crowd* term can be referred to as *controlled crowd*.

2.1 Three Genres of Crowdsourcing

In 2013, Wang et al. aggregated most of the previous studies in their very well-done survey. The mentioned work emphasizes three intuitive and well-separated genres of crowdsourcing [11]:

Games with a purpose (GWAPs), when a player without any special knowledge is put into a gaming environment and have to make right decisions to win the game under the pressure of time or any game mechanics' constraints. Phrase Detectives⁴ and JeuxDeMots⁵ can be considered as good examples of such games.

Mechanized labor (MLAB), when an annotator who meet the preliminary requirements is asked to answer a questionnaire on a centralized platform and is rewarded for their work by micropayments. The most well-recognized examples of MLAB are Amazon Mechanical Turk⁶ and CrowdFlower⁷.

Wisdom of the crowd (WOTC), when motivated volunteers share their knowledge on the given topic in the free form in order to answer some question, to explain something to other people, and so on. The obvious examples of WOTC are Wikipedia⁸ and Yahoo! Answers⁹.

Observations reveal that research papers often do not specify the exact crowdsourcing genre and treat the crowdsourcing term as a synonym to MLAB due to extreme popularity of the Amazon's product.

2.2 Cooperation in Linguistic Resources

Cooperation, derived from *to cooperate*, is to work actively with rather than against others [12, p. 435]. Unfortunately, cooperative crowdsourcing in linguistic resources is less explored in the literature, but present studies show that considering the concept of cooperation in crowdsourcing is a trending topic deserving attention.

An early study of Wikipedia and its quality by Wilkinson & Huberman [13] found a statistically significant correlation between page edits, talkpage conversations and the quality of these pages. The study revealed the fact that pages with more intense discussion activity often have better quality than less discussed ones.

⁴ <https://anawiki.essex.ac.uk/phrasedetectives/>

⁵ <http://www.jeuxdemots.org/>

⁶ <http://mturk.com/>

⁷ <http://www.crowdflower.com/>

⁸ <http://wikipedia.org/>

⁹ <https://answers.yahoo.com/>

A study by Arazy & Nov [14] pays a special attention to *local inequality* — inequality of editors' contribution in a particular article, and *global inequality* — inequality in overall Wikipedia activity for the same set of editors. As a result, they found that global inequality has an impact on local inequality, which influences editors' coordination in a positive way, which in its turn contributes to quality.

Budzise-Weaver et al. [15] consider several cases of multilingual digital libraries and their collaboration both with state institutions and crowdsourced projects in order to provide multilingual information access for users. The paper does not describe how exactly crowdsourcing can help digital libraries in doing their job, but does demonstrate significant interest to crowdsourcing from an interdisciplinary point of view.

Ranj Bar & Maheswaran [16] in their case study on Wikipedia concluded that new mechanisms are needed to coordinate the activities in crowdsourcing due to the fact that high quality articles are controlled by small groups of permanent editors, and supporting these articles is a huge burden for the editors.

3 Crowdsourcing Genres and Cooperation

Each of the three crowdsourcing genres has its own identities; and the principle of participants' cooperation changes with each particular crowdsourcing instance. However, it seems possible to denote three common points:

- **attractiveness**, the degree of how a participant can find a crowdsourcing process attractive,
- **usefulness**, the degree of how a participant can find his activity results useful to their own purposes,
- **difficulty**, the degree of how it is difficult to embed cooperative elements into a process.

When specific case studies are available, the correspondent details are provided.

3.1 Games with a purpose

The main advantage of GWAPs is their *high attractiveness*, because people love video games and it is easier to get new participants than in other genres of crowdsourcing. One may find *low usefulness* in these games, but the more attractive the game is, the less other factors are becoming important.

It is necessary to mention that video games are a very costly kind of software and producing GWAPs requires not only creating a game, but also designing innovative game mechanics allowing a player to both enjoy the game and to implicitly produce valuable data. Thus, games with a purpose have *high difficulty* to be realized.

Authors of Phrase Detectives say that the cost of data gathering using their means is lower than using other approaches [17], but they did not consider the

total cost of the game design and development. Elements of real-time players' cooperation may enhance GWAPs attractiveness even more. The evidence of this is the fact that modern cooperative multiplayer video games like Dota 2 or Destiny have substituted traditional *free for all* (deathmatch) multiplayer games.

3.2 Mechanized labor

Since MLAB projects are often deployed on specialized platforms available on the World Wide Web, the main advantage of MLAB is its *low difficulty*: cooperative elements may be embedded supplementarily to the annotation process through allowing annotators to join teams and making them participate in the team-based activity.

In order to cover as much domains as possible, platforms' owners provide only very utilitarian and generic interfaces allowing one to answer a questionnaire without exposing them to any domain-specific features.

Since MLAB participants are often rewarded for their work that may be or may not be interesting for them, the mechanized labor projects have *medium usefulness* and usually *low attractiveness*.

3.3 Wisdom of the crowd

The strong side of WOTC projects is, indeed, *high usefulness* due to the fundamental principle of such a genre, when volunteers make efforts to make their resource better for everyone. WOTC have *low attractiveness* for the same reasons, however it depends on every particular instance.

The above mentioned study by Arazy & Nov also touches upon a typical regulation problem called "edit warring" in Wikipedia [14], when "editors who disagree about the content of a page repeatedly override each other's contributions, rather than trying to resolve the disagreement through discussion".

The phenomena of "edit warring" was later studied by Yasseri et. al [18]. Such a problem may be partially resolved by using the controlled crowd instead of the public one when volunteers have a mentor and responsibility for their actions [19]. Therefore, such projects have *medium difficulty*.

4 Evidence

An evidence that cooperation does work and really stimulates participants to do more assignments is the case of OpenCorpora, which is a project focused on creation of a large annotated Russian corpus through crowdsourcing [20].

Currently, OpenCorpora participants have to annotate morphologically ambiguous examples in the MLAB manner. One can annotate examples individually, but has an opportunity to join teams and annotate examples in cooperation with their teammates. A team can be created and joined by everyone, and teams

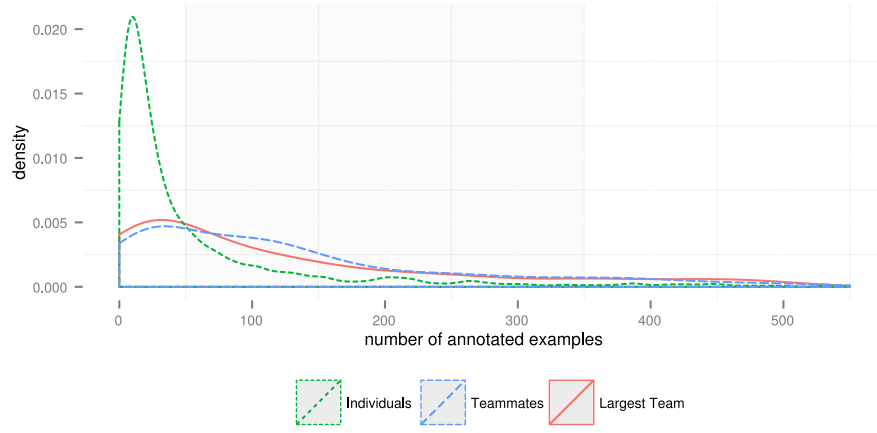


Fig. 1. Annotated examples’ densities by three user groups: the individuals, the teammates, and the largest team; the highlighted region corresponds to the interval between 50 and 350 examples.

challenge each other by means of active collaborators, annotated examples, and error rates.

As according to the full-scale pilot study conducted on one of the largest Russian information technologies’ websites¹⁰, volunteers were very positive about their participation in the cooperative annotation. The study was followed by the creation of the largest team uniting 170 participants. The team got the 2nd place in the total rank¹¹ based on the number of the annotated examples.

The possible explanation of such a result would be found in what have driven the participants’ motivation. It was not only their altruism and readiness to help, but the possibility for their team to get the leading places in the total rank, as well as their personal participation being one of the keys to the team’s possible success.

4.1 “Is there a relationship?”

To make it possible to study the present result more thoroughly, the OpenCorpora team has kindly provided us with the dataset consisted of user ID, the group’s name, and various activity information including total number of the annotated examples per user. Hereafter participants who joined a team are referred to as *teammates*, and those who did not join a team are referred to as *individuals*.

¹⁰ http://habrahabr.ru/post/152799/#comment_5315923

¹¹ <http://opencorpora.org/?page=stats>

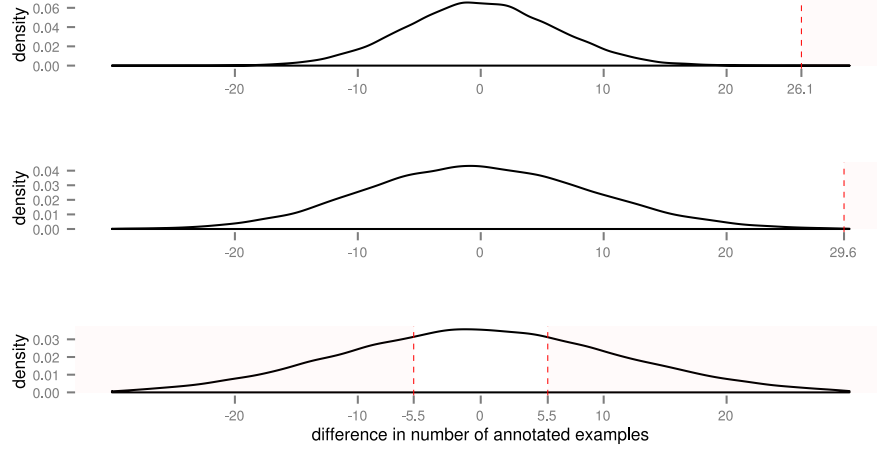


Fig. 2. Simulated differences in number of annotated examples, the vertical dashed lines represent observed differences: the upper plot (a) corresponds to H , the middle plot (b) corresponds to H' , and the lower (c) corresponds to H'' .

The initial dataset contains information on 2642 users: 2219 of them are individuals and 423 are teammates. The distributions' densities are depicted at Fig. 1 and seem to be right-skewed. In order to remove outliers from the dataset, users who annotated less than 50 examples or more than 350 examples have been excluded. As a result, the dataset has been reduced to 579 individuals and 195 teammates, 71 of those are the members of the largest team.

In general, the individuals annotated 801 531 examples and the teammates annotated 970 650, while in the dataset the individuals annotated 71 150 examples and the teammates annotated only 29 049 examples.

Hence, the research question is “*Is there a relationship between being a team member and the number of annotated examples for a regular OpenCorpora user?*”

4.2 Inference

Since the dataset is right-skewed and such hypothesis tests as *t-test* may be unreliable, a randomization test was implemented in the R programming language and executed for 25 000 times under the significance level of $\alpha = .05$ in order to estimate the unbiased *p-value*.

The true difference in means of the numbers of annotated examples among the teammates (μ_T) and the individuals (μ_I) has been examined. The following hypothesis H was evaluated in order to find a relationship between being a team member and the number of annotated examples:

$H_0 : \mu_T - \mu_I = 0$, the teammates and the individuals on average have no difference in their annotation activity,

$H_A : \mu_T - \mu_I > 0$, the teammates tend to annotate more examples on average than the individuals.

The density of differences in the number of annotated examples is demonstrated at Fig 2(a): the observed difference in means for this one-tailed test is $\bar{x}_T - \bar{x}_I = 26.085$, and the *p-value* is $p = 0$. Thus, $p < \alpha$ and the null hypothesis H_0 has been rejected, suggesting that $\mu_T > \mu_I$: the teammates tend to annotate more examples than the individuals.

5 Discussion

The obtained result can also be explained by a teammate being more loyal and attached to the resource than an individual. Therefore, it is reasonable to study the performance of a particular team.

5.1 The Largest Team vs. The Individuals

In order to compare the behavior of the individuals and the largest team members instead of all the teammates, the true difference in means of the numbers of annotated examples among the teammates of the largest team (μ_H) and the individuals (μ_I) was examined. The following hypothesis H' was evaluated in the similar way as the previous one:

$H'_0 : \mu_H - \mu_I = 0$, the teammates of the largest team and the individuals on average have no difference in their annotation activity,
 $H'_A : \mu_H - \mu_I > 0$, the teammates of the largest team annotate more examples on average than the individuals.

The simulation results for this one-tailed test are presented at Fig. 2(b): the observed difference in means is $\bar{x}_H - \bar{x}_I = 29.552$ and the *p-value* is $p = .001$. Since $p < \alpha$, the null hypothesis H'_0 has been rejected, suggesting that $\mu_H > \mu_I$: the teammates of the largest team annotate more examples than the individuals.

This result agrees well with the H_0 hypothesis and can be explained by the fact that the largest team is still relatively small and consists of only 170 teammates who were highly motivated for a short time due to news rotation on the website where they came from. Their activity decreased significantly when the announcement of the OpenCorpora disappeared from the news headline. Their team took the 2nd place on the leaderboard; they lost to the the leading team as the latter had annotated approximately seven times more examples (501 963 versus 76 559).

5.2 The Largest Team vs. Other Teams

Statistical testing of teams' performance based on comparison of their impact is complicated due to lack of participants in other teams. For instance, the 2nd largest team is comprised of 36 users only, the 3rd largest — 24, the 4th —

13, which is insufficient for any meaningful test. However, it is indeed possible to compare the performance of the largest team with the performance of other teams considered together.

The true difference in means of the numbers of annotated examples among the teammates of the largest team (μ_H) and other teams (μ_R) was examined, and the following hypothesis H'' has been evaluated:

$H_0'' : \mu_H - \mu_R = 0$, the teammates of the largest team and other teammates on average have no difference in their annotation activity,

$H_A'' : \mu_H - \mu_R \neq 0$, the teammates of the largest team and other teammates on average have the difference in their annotation activity.

The simulation results for this two-tailed test are presented at Fig. 2(c): the observed difference in means is $\bar{x}_H - \bar{x}_R = 5.453$ and the p -value is $p = .629$. Since $p > \alpha$, the null hypothesis H_0' has not been rejected, suggesting that $\mu_H = \mu_R$: the teammates of the largest team annotate the same number of examples as other teammates do.

6 Conclusion

According to the obtained results, there *is* a correlation between being a team member and the number of annotated examples for a regular OpenCorpora user. The use of team-based cooperation can improve the user activity on crowdsourced linguistic resources. However, since the study is observational, it was impossible to establish causal relationships between the variables.

When organized in teams, users do provide more annotations comparing with those who are not organized in teams. Thus, it is highly recommended for a crowdsourced resource to provide users with the opportunity to join teams and annotate examples in cooperation with their teammates.

The statistical hypotheses have been evaluated with use of the randomization test with the significance level of .05. The present dataset is available¹² in an depersonalized form under the Creative Commons Attribution-ShareAlike 3.0 license. The source code of the above mentioned simulation program is included under the MIT License.

Further work may be focused on assessing the quality of team-based cooperation results and on studying the patterns of cooperation and the efficiency of their use in other popular linguistic resources created through crowdsourcing.

Acknowledgments. This work is supported by the Russian Foundation for the Humanities, project no. 13-04-12020 “New Open Electronic Thesaurus for Russian”, and by the Program of Government of the Russian Federation 02.A03.21.0006 on 27.08.2013.

¹² <http://ustalov.imm.uran.ru/pub/opencorpora-cooperation.tar.gz>

The author would like to thank Dmitry Granovsky for the extended statistical information collected from <http://opencorpora.org/>. The author is also grateful to the anonymous referees who offered very useful comments on the present paper.

References

1. Biemann, C.: Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation* **47**(1) (2013) 97–122
2. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA) (2014) 859–866
3. Lofi, C., Selke, J., Balke, W.T.: Information Extraction Meets Crowdsourcing: A Promising Couple. *Datenbank-Spektrum* **12**(2) (2012) 109–120
4. Quinn, A.J., Bederson, B.B.: A Taxonomy of Distributed Human Computation. *Human-Computer Interaction Lab Tech Report*, University of Maryland (2009)
5. Yuen, M.C., Chen, L.J., King, I.: A Survey of Human Computation Systems. In: *International Conference on Computational Science and Engineering*, 2009. CSE '09. Volume 4., IEEE (2009) 723–728
6. Sabou, M., Bontcheva, K., Scharl, A.: Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, ACM (2012) 17:1–17:8
7. Sabou, M., Scharl, A., Michael, F.: Crowdsourced Knowledge Acquisition: Towards Hybrid-Genre Workflows. *International Journal on Semantic Web and Information Systems* **9**(3) (2013) 14–41
8. Zwass, V.: Co-Creation: Toward a Taxonomy and an Integrated Research Perspective. *International Journal of Electronic Commerce* **15**(1) (2010) 11–48
9. Erickson, T.: Some Thoughts on a Framework for Crowdsourcing. In: *CHI 2011 Workshop on Crowdsourcing and Human Computation*. (2011)
10. Suendermann, D., Pieraccini, R.: Crowdsourcing for Industrial Spoken Dialog Systems. In: Eskénazi, M., Levow, G.A., Meng, H., Parent, G., Suendermann, D., eds.: *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. John Wiley & Sons, Ltd (2013) 280–302
11. Wang, A., Hoang, C.D.V., Kan, M.Y.: Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation* **47**(1) (2013) 9–31
12. Kohn, A.: *No Contest: A Case Against Competition*. Houghton Mifflin Harcourt (1992)
13. Wilkinson, D.M., Huberman, B.A.: Cooperation and Quality in Wikipedia. In: *Proceedings of the 2007 International Symposium on Wikis*, ACM (2007) 157–164
14. Arazy, O., Nov, O.: Determinants of Wikipedia Quality: The Roles of Global and Local Contribution Inequality. In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ACM (2010) 233–236
15. Budzise-Weaver, T., Chen, J., Mitchell, M.: Collaboration and Crowdsourcing: The Cases of Multilingual Digital Libraries. *Electronic Library*, The **30**(2) (2012) 220–232

16. Ranj Bar, A., Maheswaran, M.: Case Study: Integrity of Wikipedia Articles. In: Confidentiality and Integrity in Crowdsourcing Systems. SpringerBriefs in Applied Sciences and Technology. Springer International Publishing (2014) 59–66
17. Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L.: Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Trans. Interact. Intell. Syst.* **3**(1) (2013) 3:1–3:44
18. Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertész, J.: Dynamics of Conflicts in Wikipedia. *PLOS ONE* **7**(6) (2012) e38869
19. Braslavski, P., Ustalov, D., Mukhin, M.: A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2014) 101–104
20. Bocharov, V., Alexeeva, S., Granovsky, D., Protopopova, E., Stepanova, M., Surikov, A.: Crowdsourcing morphological annotation. In: Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, RGGU (2013) 109–124